DexOPE: 6D Object Pose Estimation in Dexterous Manipulation

Ke Wu[1], Zhiwei Yang[1,2], Xiangting Meng[3], Zicheng Zhang[1],
Hui Zhang[4], Zijun Xu[1], Jieru Zhao[5], Wenchao Ding[1]*
[1]Fudan University [2]Shanghai Innovation Institute
[3]ShanghaiTech University [4]ETH Zurich
[5]Shanghai Jiao Tong University
kewu23@m.fudan.edu.cn

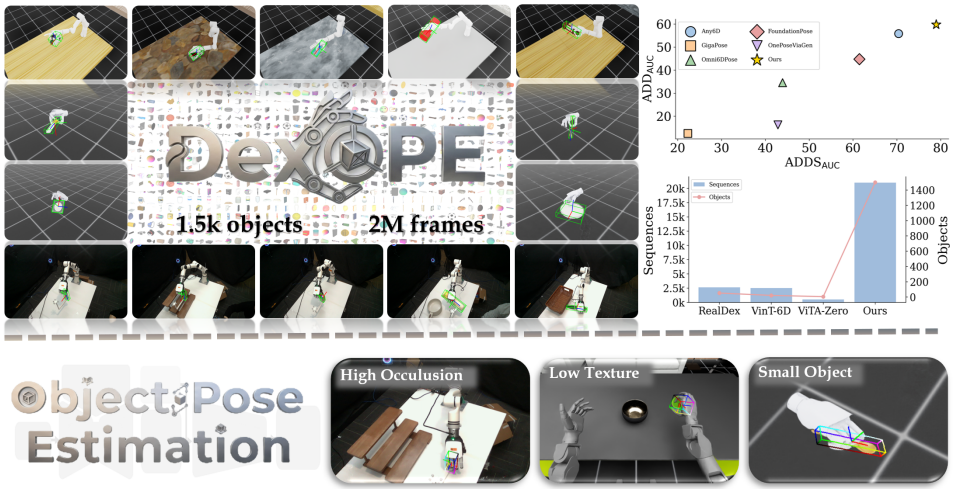

Figure 1: **Overview of the DexOPE dataset and our OPE method performance**. DexOPE contains 1.5k objects and 2M multimodal manipulation frames, significantly larger than existing datasets. Our OPE method achieves state-of-the-art ADD_{AUC} and ADDS_{AUC}, and remains robust under occlusion, low-texture surfaces, and small object geometries.

Abstract

Reliable 6D object pose estimation is essential for dexterous manipulation but remains highly challenging due to severe visual occlusions from multi-finger interactions. Progress has been limited by two key factors: the lack of large-scale manipulation datasets and the inability of existing methods to handle high uncertainty under occlusion. To address these challenges, we introduce DexOPE, a large-scale dataset and a physics-guided pose estimation framework. DexOPE provides a multimodal dataset with 1.5k diverse objects, an order of magnitude larger than prior work, covering simulated and real-world grasping and manipulation. To handle pose ambiguity under occlusion, we propose a Physics-Guided Score Matching approach that models the pose posterior using score-based diffusion models rather than deterministic regression. Physical constraints from hand–object interactions, including nonpenetration and tactile contact, are incorporated as guidance during sampling, enabling physically plausible pose inference even under severe or complete occlusion. Extensive experiments show that DexOPE

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

achieves state-of-the-art performance, significantly outperforming existing methods in robustness to occlusion and generalization to unseen objects.

1 Introduction

Enabling robots to achieve human-like dexterous hand manipulation is one of the ultimate goals of embodied intelligence. In this process, high-frequency and accurate 6D object pose estimation (OPE) lies at the core of the perception–control loop. Unlike fixed end-effector grasping, dexterous manipulation is inherently dynamic: frequent multi-finger interactions induce persistent and severe visual occlusions, posing significant challenges to object pose estimation.

Under severe or complete occlusion, visual observations alone are insufficient for accurate pose estimation, making tactile sensing essential. Existing object pose estimation (OPE) methods can be broadly categorized by input modality. Vision-dominant approaches Wen et al. [2024], Lee et al. [2025], Zhang et al. [2024b] rely on feature matching, regression, or diffusion models, but fail when objects are heavily occluded by fingers due to the loss of visual cues. Recent work has begun incorporating tactile information into OPE Wan et al. [2024], Li et al. [2025a], yet most such methods depend on test-time optimization, which is susceptible to local minima. Moreover, they are typically limited to fewer than 50 objects and struggle to generalize to real-world dynamic interactions. This limitation largely arises from the high cost of collecting large-scale dexterous manipulation data, which is substantially more expensive than standard perception datasets.

To address these challenges, as shown in Figure 1, we present DexOPE, a unified framework that combines a large-scale dexterous manipulation dataset with a physics-aware object pose estimation method. On the data side, we design a scalable data engine based on RL, imitation learning, and teleoperation, covering four types of hand–object interactions in both simulation and the real world. This enables us to significantly scale up data collection and construct a large-scale dexterous manipulation dataset comprising 1.5k object categories and 2M frames, more than an order of magnitude larger than existing datasets.

On the method side, we formulate object pose estimation as a temporal relative pose estimation problem and propose a physics-guided visual–tactile diffusion framework. Interaction-aware physical constraints are incorporated as gradient guidance during diffusion sampling, effectively compensating for missing visual features under severe occlusion. Temporal pose estimation further resolves multi-hypothesis ambiguities and enforces physical consistency, resulting in robust and reliable pose estimation.

Our contributions are summarized as follows:

- We introduce a scalable data engine for dexterous grasping and manipulation, and construct a large-scale dataset with 1.5k objects and 2M frames, substantially surpassing prior work in scale and diversity.
- We propose a physics-guided visual–tactile diffusion framework for object pose estimation, where vision provides a global prior while proprioception and tactile feedback serve as dynamic local constraints, enabling accurate pose inference even under severe or complete visual occlusion.
- Extensive experiments demonstrate that DexOPE significantly outperforms state-of-the-art vision-based methods under heavy occlusion, low-texture conditions, and achieves strong reliability in complex real-world dexterous manipulation tasks.

2 Related Works

2.1 Object Pose Estimation

Model-free 6D object poses estimation remains a major challenge in robotic perception. Recent work has explored diverse strategies. Template-based methods such as GigaPose Nguyen et al. [2024] enable fast and robust pose estimation using discriminative templates and patch correspondences, while FoundationPose Wen et al. [2024] bridges model-based and model-free settings by

Table 1: Comparison of our dataset with existing dexterous manipulation datasets. Our dataset uniquely supports both Grasping and Manipulation in Simulation and Real-world environments, with a significantly higher number of objects and sequences.

Dataset	Modality			Action Mode		Objs.	Seqs.	Frames
	RGB	Depth	Contact F.	Grasp	Manip.			
RealDex	✓	✓	✗	✓	✗	52	2.63k	59k
VinT6D-Sim	✓	✓	✓	✓	✗	21	2.5k	2M
VinT6D-Real	✓	✓	✓	✓	✗	21	0.1k	100k
ViTa-Zero	✓	✓	✓	✓	✗	5	-	-
Ours								
SIM GRASP	✓	✓	✓	✓	✗	1.5k	20k	2M
SIM MANIP	✓	✓	✓	✓	✓		1k	300k
REAL GRASP	✓	✓	✓	✓	✗		0.4k	40k
REAL MANIP	✓	✓	✓	✓	✓		0.2k	20k

leveraging neural implicit representations for novel view synthesis when CAD models are unavailable, showing strong generalization to unseen objects. Model-free approaches further reduce input requirements: Any6D Lee et al. [2025] estimates both pose and scale from a single RGB-D anchor image. Universal frameworks have also emerged, with Omni6DPose Zhang et al. [2024b] introducing a benchmark and model for universal 6D pose estimation and tracking to address the lack of large-scale datasets. Despite these advances, challenges persist for objects with low texture, high symmetry, and severe occlusion. In one-shot settings, OnePoseViaGen Geng et al. [2025] tackles arbitrary novel objects via single-view 3D generation and domain randomization. However, vision-centric methods often struggle to maintain physical plausibility under the extreme occlusions typical of dexterous manipulation. In contrast, DexOPE explicitly incorporates interaction physics to resolve these ambiguities.

2.2 Object Pose Estimation and Manipulation Datasets

High-quality, diverse benchmarks are fundamental to robust pose estimation. While the BOP Challenge Hodan et al. [2018] standardized evaluation across varying difficulties, and traditional sets like LM-O and TOYL addressed heavy occlusion and lighting, recent work targets more complex scenarios. DynOPETs Meng et al. [2025] introduces dynamic objects with moving cameras, while VinT-6D Wan et al. [2024] provides a multimodal (vision, touch, proprioception) sim-to-real dataset for in-hand manipulation within the "Perception-Planning-Control" paradigm. Additionally, ViTa-Zero Li et al. [2025a] explores zero-shot visuotactile estimation for better generalization, and GraspClutter6D Back et al. [2025] focuses on cluttered grasping. Despite these advances, existing datasets remain limited in scale (<100 objects) and focus on static states, overlooking the continuous contact dynamics of manipulation. To bridge this gap, we introduce DexOPE, comprising 1.5k objects, an order of magnitude larger than prior work, and spanning four sim-to-real domains to enable robust category-level generalization. Furthermore, establishing a scalable data generation pipeline is critical for overcoming the scarcity of dexterous manipulation data Liu et al. [2026], Shen et al. [2026], Li et al. [2026a].

3 Method

We will introduce our data engine and pose estimation method in this section.

3.1 DexOPE Dataset

3.1.1 SIM GRASP Generation

Unlike parallel-jaw grippers that execute open-loop grasps based on geometric heuristics or analytic force-closure metrics Fang et al. [2023], generating trajectories for multi-fingered dexterous hands involves high-dimensional action spaces and complex contact dynamics. To scale data generation across a diverse range of objects, as shown in Figure 2, we employ an RL approach.

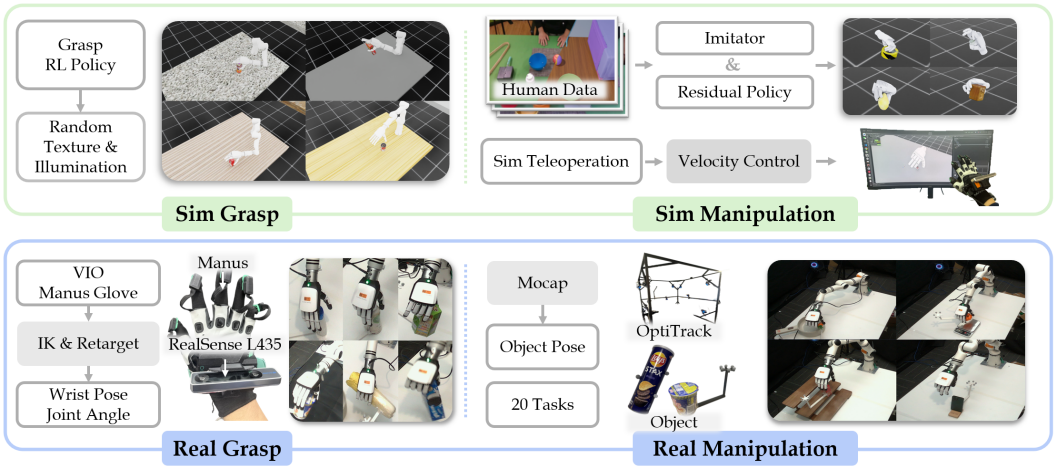

Figure 2: **DexOPE Data Engine**. The DexOPE data generation framework spans four domains : (1) Sim Grasp Employs an Objective-Driven RL policy to generate diverse, stable grasps across 1.5k objects. (2) Sim Manipulation: Combines a trajectory imitator with a residual policy to retarget human demonstrations into robotic manipulation sequences. (3) Real Grasp: Captures naturalistic grasping data via a MANUS VR glove and teleoperation. (4) Real Manipulation: Utilizes a custom marker-based tracking system and OptiTrack to provide precise 6D pose ground truth even under severe multi-finger occlusion.

Object Assets and Layout Generation. We curated a dataset of 1,429 objects, including 73 items from YCB Calli et al. [2015], AnyGrasp Fang et al. [2023] and 1,356 objects from Objaverse Deitke et al. [2023]. At the beginning of each episode, object poses are randomized initialized within the reachable workspace of the robot.

Grasp Synthesis and Trajectory Generation. To efficiently produce large-scale dexterous manipulation trajectories, we utilize an RL-based framework. The observation space consists of a robust, hand-centric representation Zhang et al. [2024a, 2025]: the nearest distance vectors from finger joints to the object surface, along with the joint angles. The reward function prioritizes hand-object proximity, successful grasping, and velocity regularization. A grasp is deemed successful only if the object is lifted to a target height and remains stable without slipping for a fixed duration. Our RL policy achieves an 89% success rate across 1,429 objects, demonstrating exceptional generalization and robustness.

3.1.2 SIM MANIP Generation

Sim Grasp mainly targets data generation for simple grasping actions, whereas dexterous manipulation involves continuous state transitions and complex contact dynamics. Collecting such data via teleoperation is both labor-intensive and difficult to scale. To address this challenge, we leverage large-scale human demonstration data. Building on ManipTrans Li et al. [2025c], we implement a two-stage RL framework consisting of a generalist trajectory imitator and a physics-aware residual policy. This framework decouples human motion imitation from physical interaction correction, enabling scalable generation of high-fidelity and physically plausible manipulation trajectories.

Under this unified framework, we adapt a subset of the OakInk2 Zhan et al. [2024] dataset to construct the Sim Manip dataset, which contains 209 tasks, including screw, write, and pour—with a total of 999 trajectories. In each simulation episode, the pre-trained residual policy is executed to track the human reference manipulation. We apply strict success criteria and retain only episodes without physical violations, such as severe interpenetration or object dropping. For successful episodes, we record dense trajectory data, including object 6D poses, hand joint angles, and interaction states. To support robust visual learning, we replay these trajectories in IsaacSim and randomize camera viewpoints. In addition, we employ teleoperation in IsaacLab to handle particularly challenging tasks.

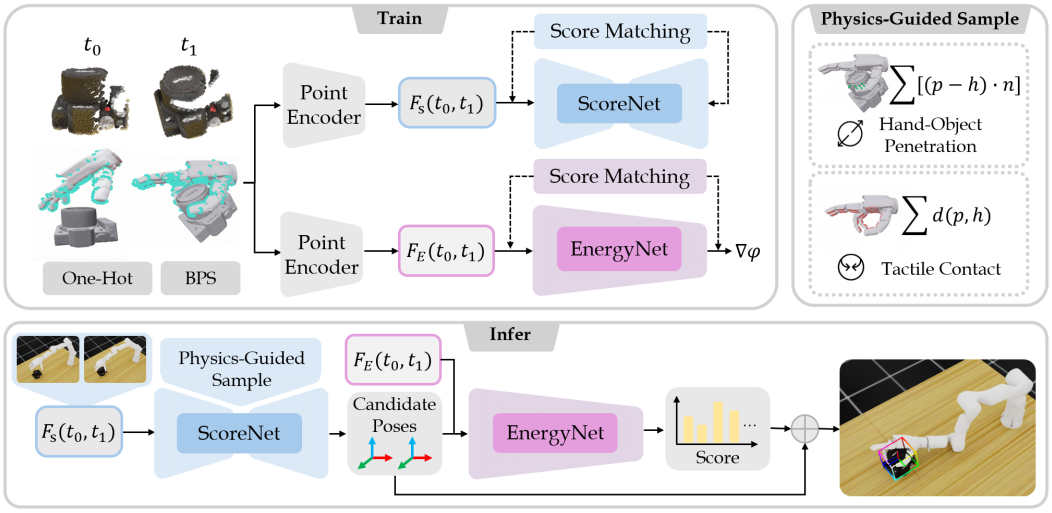

Figure 3: **Physics-Guided Score-Matching OPE Method Pipeline**. The framework takes object point clouds at adjacent time steps as input. It employs score based diffusion conditioned on multimodal point cloud features and shape priors to model the pose posterior. Physical constraints including non penetration and tactile contact provide gradient guidance during sampling to resolve occlusions. Final predictions are selected using EnergyNet ranking and clustering to ensure physically plausible results.

3.1.3 REAL GRASP and REAL MANIP Acquisition

Acquiring high-fidelity 6D pose annotations during dexterous manipulation is challenging because intense hand-object interactions cause severe visual occlusion and frequently dislodge surface-mounted optical markers. To bridge the sim-to-real gap, we developed a robust acquisition pipeline combining a custom marker-based tracking system with a safe teleoperation framework.

To ensure reliable ground truth even under full occlusion (Figure 6), we use a 3D-printed attachment with five MoCap markers fixed to the object's non-interacting end. We initialize the pose using FoundationPose Wen et al. [2024] on a static, unoccluded frame to compute the constant transformation between the object and the marker rigid body. This allows us to recover the 6D pose for all subsequent frames solely via markers, effectively decoupling pose estimation from both visual occlusion and physical marker detachment.

For data generation, we use a MANUS VR glove to retarget human motions to the robot via Inverse Kinematics (IK), with real-time collision detection to ensure hardware safety. The final phase records time-synchronized RGB-D, MoCap, proprioception, and tactile data. The resulting dataset features two subsets: REAL GRASP (10 small, low-texture objects) and REAL MANIP (10 functional tasks, from precision placement to force-sensitive weighing).

3.2 Object Pose Estimation Method

3.2.1 Problem Formulation and Input Representation

We formulate object pose estimation during manipulation as a Partially Observable Markov Process (POMP) Li et al. [2026b,c]. Although the object is typically visible before manipulation begins (during the approach phase), severe finger occlusions in dexterous manipulation make direct frame-by-frame regression of the absolute pose extremely challenging.

To address this challenge, we abandon direct prediction of absolute 6D poses and instead estimate the relative pose change between consecutive time steps, denoted as $\hat{T}^t_{t-1}$. Absolute poses depend on specific coordinate frame definitions and are inherently ambiguous, whereas relative motions induced by manipulation (e.g., "the object is pushed left by a finger") are physically consistent and invariant to the choice of coordinate system. This makes them easier for the network to learn and generalize.

We aim to design a data generation and object pose estimation method suitable for hand-object interaction in dexterous manipulation. Using the dexterous hand's point cloud is the most universal representation. The model takes as input object point clouds at two consecutive time steps, $\{P_{t-1}, P_t\}$, along with the corresponding dexterous hand point clouds $\{H_{t-1}, H_t\}$, generated via forward kinematics (FK). To enhance the network's geometric perception and distinguish between categories, we additionally incorporate one-hot category encodings and Basis Point Set (BPS) Prokudin et al. [2019] global shape features B_0 derived from the initial complete point cloud P_0.

To mitigate distribution shift from rapid arm motions, we propose a translation–rotation decoupled normalization: translational components are expressed in a wrist-centered local frame while rotations remain in the world frame. This ensures that normalized point clouds stay unchanged when hand and object move together, while world-frame rotations preserve gravity direction for learning gravity-induced dynamics. Inference efficiency is further improved by activating diffusion-based prediction only when FK-computed hand keypoints enter a predefined neighborhood around the object.

3.2.2 Score-Based Pose Generation

To address the multi-hypothesis ambiguity caused by object symmetry and severe occlusion, inspired by previous diffusion-based methods Zhang et al. [2024b], we model relative pose $\hat{T}_{t-1}^{t}$ estimation as a score-based conditional diffusion process. Our goal is to learn the conditional probability distribution $p(T|\mathcal{O})$ of the relative pose transformation, from which we sample K high-confidence pose hypotheses. This generative paradigm enables the network to retain multiple plausible motion possibilities even under highly ambiguous observations, thereby avoiding convergence to an erroneous mean state.

As shown in Figure 3, the input to ScoreNet, denoted as $\mathcal{O}$, consists of the object point clouds at consecutive time steps $\{\tilde{P}_{t-1}, P_t\}$, the corresponding hand point clouds $\{H_{t-1}, H_t\}$, a one-hot encoding, and the BPS encoding B_0. All inputs are processed by a PointNet++ Qi et al. [2017] based encoder to extract geometric features, which are then fused with the time step t (mapped via an MLP) to serve as the conditional input for the denoising network. To mitigate error accumulation in continuous manipulation and to handle extreme cases, such as missing depth observations due to complete occlusion at time $t-1$, we adopt a ground-truth re-alignment training strategy. During training, instead of using partial point clouds directly from sensors, we transform the initial complete point cloud P_0 to the previous state using the ground truth pose T_{gt}^{t-1}, denoted as $\tilde{P}_{t-1} = T_{gt}^{t-1} \cdot P_0$. This strategy compels the network to learn how to infer dynamic changes from current partial observations given a deterministic history, ensuring the model can perform reasonable deductions based on historical priors even when visual data is lost.

For the diffusion implementation, we utilize the Variance Exploding (VE) SDE Song et al. [2020] to construct the denoising process. The forward process gradually blurs the relative pose distribution by injecting Gaussian noise, while the reverse generative process estimates the score function of the perturbed data by training a time-dependent Score Network $s_\theta(T, t, \mathcal{O})$. We employ Denoising Score Matching Vincent [2011] for training, with the objective function defined as:

$$\mathcal{L}_{score} = \mathbb{E}_{t,T,\mathbf{z}} \left[\lambda(t) \left\| s_\theta(T + \sigma(t)\mathbf{z}, t, \mathcal{O}) + \frac{\mathbf{z}}{\sigma(t)} \right\|_2^2 \right] \tag{1}$$

where $\sigma(t)$ represents the noise scale function and $\lambda(t)$ is the weighting coefficient.

3.2.3 Physics-Guided Sampling Strategy

Once the Score Network is trained, we sample candidate poses from the prior distribution by solving the Probability Flow ODE Song et al. [2020]. However, poses generated solely from visual data often lack physical plausibility.

To enforce physical plausibility, we incorporate a gradient-based noise correction mechanism into the sampling process, using differentiable physical constraints to guide pose generation. At each sampling step, the initial object point cloud P_0 is transformed by the current estimated pose $\hat{T}_0^t$ to obtain $\hat{P}_t = \hat{T}_0^t \cdot P_0$. Given $\hat{P}_t$ and the hand point cloud H_t, we define two complementary energy terms: a penetration energy and a tactile attraction energy.

For penetration handling, we introduce a normal-aligned repulsion energy $\mathcal{L}_{Pen}$. For each hand point h_t, we find its nearest object point p_t and surface normal n, and penalize penetration detected via the signed projection below zero:

$$\mathcal{L}_{Pen} = \frac{1}{N} \sum_{h \in H_t} \max\left(0, (p - h) \cdot n\right) \tag{2}$$

A positive dot product indicates that the hand point lies "inside" the object surface (against the normal direction), generating a repulsion gradient; otherwise, the loss is zero. This design leverages surface normal information, handling complex geometric intersections more robustly than simple distance thresholding.

To exploit tactile feedback, we define a tactile contact energy $\mathcal{L}_{Tac}$ that pulls the object surface toward activated sensor regions. Specifically, for active tactile points h_{active}, we apply a truncated one-way Chamfer distance between sensor points and object points.

$$\mathcal{L}_{Tac} = \frac{1}{|h_{active}| + \epsilon} \sum_{p_s \in h^*_{active}} \min_{p \in \hat{P}_{obj}} \|p_s - p\|_2 \tag{3}$$

This energy uses tactile signals as a strong space prior to "pull" the object towards the fingers.

Finally, instead of directly applying gradients as in Langevin Dynamics Welling and Teh [2011], we adopt a stable noise-mixing strategy. At each denoising step t, we compute the gradient of the total physical energy with respect to T and form a guidance direction along its descent. The final update direction is obtained by interpolating between the original diffusion sampling direction and the physically guided direction, controlled by a guidance scale α.

$$\begin{aligned} d_{mix} &= d_{sample} + \alpha \cdot (d^* - d_{sample}) \\ T_{t-1} &= T_{mean} + d_{mix} \end{aligned} \tag{4}$$

This rectifies the diffusion trajectory in noise space, producing poses that are consistent with visual observations while satisfying physical interaction constraints.

3.2.4 Energy-Based Ranking and Inference

While score-based diffusion models are capable of generating diverse pose candidates, an efficient mechanism is required to assess their quality during the inference phase. Since calculating exact likelihoods via direct integration of the Probability Flow ODE Song et al. [2020] is prohibitively time-consuming for real-time operation, we train an auxiliary EnergyNet Ψ_ϕ to efficiently approximate the log-likelihood. This network shares the same architecture as the ScoreNet and is trained via Fisher Divergence Matching, aligning its gradient with the score function:

$$\mathcal{L}_{energy} = \mathbb{E}_{t \sim \mathcal{U}(\epsilon, 1)} \left[\lambda(t) \left\| \nabla_T \Psi_\phi(T_t, t | \mathcal{O}) + \frac{T_0 - T_t}{\sigma_t^2} \right\|_2^2 \right] \tag{5}$$

where $\frac{T_0 - T_t}{\sigma_t^2}$ represents the target score derived from the perturbation kernel. During inference, we rank the generated candidate poses based on likelihoods output by the EnergyNet and discard the bottom 40% of low-confidence samples to eliminate outliers. Crucially, because occlusion and symmetry can induce multimodal distributions, simply averaging the remaining candidates would result in erroneous predictions (the mean-mode problem). Therefore, we employ the DBSCAN Ester et al. [1996] clustering algorithm to identify dominant peaks within the pose distribution and select the densest cluster as the final prediction mode, effectively resolving multi-solution ambiguity. Finally, we aggregate the optimal pose within the selected cluster. For the translation component $\hat{t}$, we apply standard mean pooling; for the rotation component, we compute the optimal quaternion $\hat{q}$ by solving for the maximum eigenvector Markley et al. [2007].

4 Experiments

We present comparative results across different datasets and we conduct ablation studies to evaluate each stage of our framework. The real world setup is shown in Figure 4. Additional details are provided in the supplementary material.



Figure 4: **Real-world experiment setup**. Left: robotic arm (Realman RM65B) equipped with an InspireFTP dexterous hand and RealSense L515 camera. Right: 20 test objects.

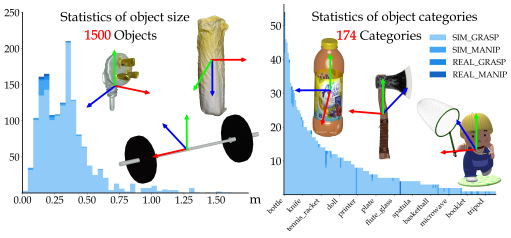

Figure 5: **Statistics of objects in DexOPE Dataset**. Left: Object size distribution across 1500 objects and diverse shapes. Right: Category distribution, highlighting 174 categories.

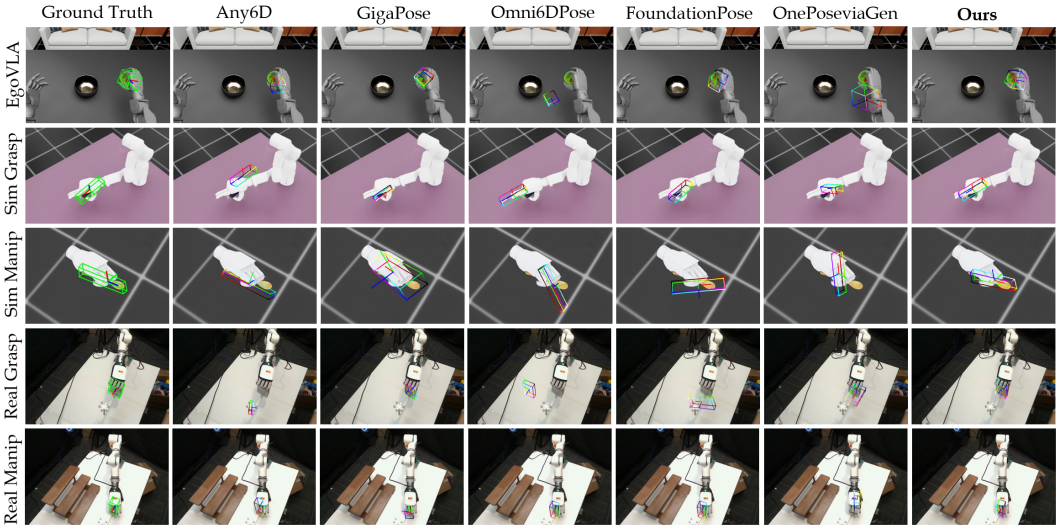

Figure 6: **Qualitative Object Pose Estimation Results**. We conducted comparative experiments using the public hand-object manipulation dataset EgoVLA and our own DexOPE dataset. Our method was evaluated against two model-based methods Wen et al. [2024], Nguyen et al. [2024] and four model-free methods Lee et al. [2025], Zhang et al. [2024b], Geng et al. [2025], Li et al. [2025b]. The results demonstrate that our approach significantly outperforms existing methods in both simulation and real-world environments.

4.1 Datasets and Metrics

Datasets. Our DexOPE dataset constitutes a large-scale benchmark comprising 2M frames and 1.5k unique objects collected from four distinct sources: SIM GRASP, SIM MANIP, REAL GRASP, and REAL MANIP. The comprehensive statistics of object categories and sizes are both demonstrated in Figure 5. We additionally evaluate our method and baselines on the EgoVLA Yang et al. [2025] dataset, which focuses on egocentric manipulation with complex dynamics and severe occlusions.

Metrics We evaluate pose accuracy using the Average Distance of Model Points (ADD) Deng et al. [2024], applying ADD for asymmetric objects and ADD-S for symmetric ones, and report the corresponding AUC scores Xiang et al. [2017]. We also assess 3D bounding box accuracy using $\text{IoU}_{25,50,75}$ Zhang et al. [2024b].

4.2 Comparison on DexOPE Dataset.

We compare our approach on DexOPE Dataset. Despite the inherent prior advantage of model-based methods, our quantitative results (As shown in Table 2) demonstrates that our method achieves superior performance across both simulation and real-world domains. In Sim Grasp scenarios, our method achieves the lowest pose errors. Crucially, in the challenging Sim Manip dynamic sequences, our method outperforms all baselines by a substantial margin, achieving ADD_{AUC}= 57.58

and ADDS_{AUC}= 74.14, effectively resolving pose ambiguities during continuous object reorientation. The superiority of our method extends to the real-world domain. In Real Grasp, we achieve state-of-the-art performance. Furthermore, in Real Manip tasks characterized by complex contact dynamics, our method demonstrates exceptional robustness, recording the lowest ADD_{Err} of 0.0334 and the highest ADD_{AUC} of 65.56.

Table 2: Quantitative comparison on SIM and REAL datasets. Results are reported separately for SIM (top) and REAL (bottom). Best results are in bold.

Method	**SIM GRASP**				**SIM MANIP**			
	$\text{ADD}_{\text{Err}} \downarrow$	$\text{ADDS}_{\text{Err}} \downarrow$	$\text{ADD}_{\text{AUC}} \uparrow$	$\text{ADDS}_{\text{AUC}} \uparrow$	$\text{ADD}_{\text{Err}} \downarrow$	$\text{ADDS}_{\text{Err}} \downarrow$	$\text{ADD}_{\text{AUC}} \uparrow$	$\text{ADDS}_{\text{AUC}} \uparrow$
Any6D	0.0954	0.0710	**55.55**	72.45	0.1912	0.1651	51.64	58.00
GigaPose	0.1357	0.1051	20.24	33.81	0.2674	0.2235	10.42	17.61
Omni6DPose	0.1771	0.1406	14.52	25.36	0.2315	0.1952	23.95	31.54
Foundation Pose	0.0884	0.0618	54.06	62.63	0.1642	0.1220	24.37	41.72
OnePoseviaGen	0.1330	0.0872	17.38	45.42	0.2546	0.2087	14.89	37.58
V-HOP	0.1039	0.0464	14.34	55.46	0.0627	0.0415	37.27	58.59
Ours	**0.0501**	**0.0248**	49.81	**75.23**	**0.0434**	**0.0259**	**57.58**	**74.14**
Method	**REAL GRASP**				**REAL MANIP**			
	$\text{ADD}_{\text{Err}} \downarrow$	$\text{ADDS}_{\text{Err}} \downarrow$	$\text{ADD}_{\text{AUC}} \uparrow$	$\text{ADDS}_{\text{AUC}} \uparrow$	$\text{ADD}_{\text{Err}} \downarrow$	$\text{ADDS}_{\text{Err}} \downarrow$	$\text{ADD}_{\text{AUC}} \uparrow$	$\text{ADDS}_{\text{AUC}} \uparrow$
Any6D	0.1324	0.1020	55.59	76.58	0.1276	0.1037	60.58	74.56
GigaPose	0.3466	0.2809	9.01	20.29	0.4231	0.3487	10.34	17.90
Omni6DPose	0.0926	0.0669	49.96	60.96	0.1384	0.1116	49.77	56.69
Foundation Pose	0.1001	0.0592	40.14	66.82	0.1075	0.0393	28.92	73.63
OnePoseviaGen	0.2322	0.1750	14.00	43.26	0.1520	0.0987	18.93	45.28
Ours	**0.0342**	**0.0156**	**65.87**	**84.44**	**0.0334**	**0.0171**	**65.56**	**81.51**

4.3 Comparison on EgoVLA.

We evaluate on the EgoVLA dataset (egocentric manipulation with severe hand occlusion and rapid motion). As shown in Table 3, our method achieves state-of-the-art performance across all metrics, outperforming strong baselines including Omni6DPose and FoundationPose, with the lowest pose errors (ADD_{Err}= 0.0287 , ADDS_{Err}= 0.0138) and highest bounding box accuracy $\text{IoU}_{25}(\%)$ = 41.29 and $\text{IoU}_{50}(\%)$ = 30.12, demonstrating that incorporating interaction physics is critical for accurate and physically plausible pose estimation.

Table 3: Quantitative comparison with state-of-the-art methods. The best results are in bold.

Method	$\text{ADD}_{\text{Err}} \downarrow$	$\text{ADDS}_{\text{Err}} \downarrow$	$\text{ADD}_{\text{AUC}} \uparrow$	$\text{ADDS}_{\text{AUC}} \uparrow$	$\text{IoU}_{25}(\%) \uparrow$	$\text{IoU}_{50}(\%) \uparrow$	$\text{IoU}_{75}(\%) \uparrow$
Any6D	0.1519	0.0931	6.44	46.86	10.71	1.36	0.00
GigaPose	0.1613	0.1236	23.65	50.82	19.90	9.92	2.27
Omni6DPose	0.1168	0.0922	57.98	61.29	18.75	11.13	4.90
FoundationPose	0.0629	0.0285	50.47	71.51	37.24	26.59	1.57
OnePoseviaGen	0.1287	0.0776	13.43	45.93	17.82	7.38	0.99
Ours	**0.0287**	**0.0138**	**60.04**	**77.50**	**41.29**	**30.12**	**7.27**

4.4 Ablation Study

Table 4: Ablation Study on the Physics-Guided Sampler.

Pen.	**Cont.**	$\text{ADD}_{\text{Err}} \downarrow$	$\text{ADDS}_{\text{Err}} \downarrow$	$\text{ADD}_{\text{AUC}} \uparrow$	$\text{ADDS}_{\text{AUC}} \uparrow$
✗	✗	0.0613	0.0342	40.77	67.76
✗	✓	0.0584	0.0311	41.75	69.14
✓	✗	0.0576	0.0304	42.41	69.61
✓	✓	**0.0501**	**0.0259**	**49.89**	**74.43**

Table 5: Ablation on γ and K.

	Value	$\text{ADD}_{\text{Err}} \downarrow$	$\text{ADDS}_{\text{Err}} \downarrow$	$\text{ADD}_{\text{AUC}} \uparrow$	$\text{ADDS}_{\text{AUC}} \uparrow$
γ	1.0	0.0540	0.0298	46.01	70.20
	0.8	0.0529	0.0286	47.15	71.41
	0.4	0.0513	0.0260	49.05	74.16
	0.2	**0.0501**	**0.0259**	**49.89**	**74.43**
K	20	0.0514	0.0270	48.62	72.98
	40	0.0509	0.0266	49.10	74.45
	100	0.0500	0.0260	49.95	74.47
	200	**0.0499**	**0.0258**	**50.13**	**74.56**

4.4.1 Ablation on Physics-Guided Sampler

To validate the effectiveness of our physics-guided sampler, we conducted an ablation study on 30 scenes selected from Sim-Grasp and reported the averaged results. As shown in Table 4, incorporating physical guidance from hand-object interactions significantly improves pose estimation accuracy, particularly in sequences characterized by severe occlusion Shen et al. [2026].

4.4.2 Ablation on Candidate Poses

We ablate the filtering ratio γ and candidate number K in Table 5. Low-likelihood poses degrade results even after clustering, and performance saturates around $K{=}50$. we thus set $K{=}50$ to balance accuracy and memory cost.

5 Conclusion and Limitations

DexOPE introduces a large-scale multimodal dataset along with a physics-guided diffusion approach that enforces physical constraints to produce plausible pose estimates under severe occlusion. The framework requires an initial unoccluded frame to obtain the object point cloud prior P_0, and inference speed, while sufficient for robot operations, falls short of real-time performance; we plan to accelerate this via TensorRT.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62403142, and in part by the Science and Technology Commission of Shanghai Municipality under Grant 24511103100.

References

Seunghyeok Back, Joosoon Lee, Kangmin Kim, Heeseon Rho, Geonhyup Lee, Raeyoung Kang, Sangbeom Lee, Sangjun Noh, Youngjin Lee, Taeyeop Lee, et al. Graspclutter6d: A large-scale real-world dataset for robust perception and grasping in cluttered scenes. *arXiv preprint arXiv:2504.06866*, 2025.

Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, pages 510–517. IEEE, 2015.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023.

Yu Deng, Jiahong Xue, Teng Cao, Yingxing Zhang, Lanxi Wen, and Yiyang Chen. Superpose: Improved 6d pose estimation with robust tracking and mask-free initialization. *arXiv preprint arXiv:2409.19986*, 2024.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics (T-RO)*, 2023.

Zheng Geng, Nan Wang, Shaocong Xu, Chongjie Ye, Bohan Li, Zhaoxi Chen, Sida Peng, and Hao Zhao. One view, many worlds: Single-image to 3d object meets generative domain randomization for one-shot 6d pose estimation. *arXiv preprint arXiv:2509.07978*, 2025.

Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018.

Yiming Jiang, Chengzhen Yan, Qichao Ma, Qingchen Liu, and Jiahu Qin. Balance: Body-aware locomotion with adaptive networked critics and estimation for conflicting objectives on legged robots. *IEEE Transactions on Industrial Electronics*, 2026.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Taeyeop Lee, Bowen Wen, Minjun Kang, Gyuree Kang, In So Kweon, and Kuk-Jin Yoon. Any6d: Model-free 6d pose estimation of novel objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11633–11643, 2025.

Boyi Li, Yifan Shen, Yuanzhe Liu, Yifan Xu, Jiateng Liu, Xinzhuo Li, Zhengyuan Li, Jingyuan Zhu, Yunhan Zhong, Fangzhou Lan, et al. Toward cognitive supersensing in multimodal large language model. *arXiv preprint arXiv:2602.01541*, 2026a.

Hongyu Li, James Akl, Srinath Sridhar, Tye Brady, and Taşkın Padır. Vita-zero: Zero-shot visuotactile object 6d pose estimation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16050–16057. IEEE, 2025a.

Hongyu Li, Mingxi Jia, Tuluhan Akbulut, Yu Xiang, George Konidaris, and Srinath Sridhar. V-hop: Visuo-haptic 6d object pose tracking. *arXiv preprint arXiv:2502.17434*, 2025b.

Jialuo Li, Hai-Tao Zhang, Ran Shi, Bin Liu, and Yang Tang. Adaptive gaussian lidar-inertial slam for autonomous surface vehicles operating in degraded aquatic scenarios. *IEEE Transactions on Industrial Electronics*, 2026b.

Jialuo Li, Hai-Tao Zhang, Ran Shi, Jiayu Zou, Bin Liu, and Yang Tang. Lidar-inertial slam with voxel-information filter for autonomous surface vehicles. *IEEE Transactions on Industrial Electronics*, 2026c.

Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6991–7003, 2025c.

Yuanzhe Liu, Jingyuan Zhu, Yuchen Mo, Gen Li, Xu Cao, Jin Jin, Yifan Shen, Zhengyuan Li, Tianjiao Yu, Wenzhen Yuan, et al. Palm: Progress-aware policy learning via affordance reasoning for long-horizon robotic manipulation. *arXiv preprint arXiv:2601.07060*, 2026.

F Landis Markley, Yang Cheng, John L Crassidis, and Yaakov Oshman. Averaging quaternions. *Journal of Guidance, Control, and Dynamics*, 30(4):1193–1197, 2007.

Xiangting Meng, Jiaqi Yang, Mingshu Chen, Chenxin Yan, Yujiao Shi, Wenchao Ding, and Laurent Kneip. Dynopets: A versatile benchmark for dynamic object pose estimation and tracking in moving camera scenarios. *IEEE Robotics and Automation Letters*, 10(10):10298–10305, 2025. doi: 10.1109/LRA.2025.3597857.

Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. Gigapose: Fast and robust novel object pose estimation via one correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9903–9913, 2024.

Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4332–4341, 2019.

Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017.

Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in vlms. *Advances in Neural Information Processing Systems*, 38:17929–17960, 2026.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Zhaoliang Wan, Yonggen Ling, Senlin Yi, Lu Qi, Wangwei Lee, Minglei Lu, Sicheng Yang, Xiao Teng, Peng Lu, Xu Yang, et al. Vint-6d: A large-scale object-in-hand dataset from vision, touch and proprioception. *arXiv preprint arXiv:2501.00510*, 2024.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024.

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 445–456, 2024.

Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. Graspxl: Generating grasping motions for diverse objects at scale. In *European Conference on Computer Vision*, pages 386–403. Springer, 2024a.

Hui Zhang, Zijian Wu, Linyi Huang, Sammy Christen, and Jie Song. Robustdexgrasp: Robust dexterous grasping of general objects. *arXiv preprint arXiv:2504.05287*, 2025.

Jiyao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, and Hao Dong. Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In *European Conference on Computer Vision*, pages 199–216. Springer, 2024b.

A Implementation Details

A.1 Architecture of the Score Network

The ScoreNet architecture utilizes a modular design to extract and fuse multimodal features for pose estimation. A PointNet++ backbone encodes geometric relationships between the hand and object, while a Gaussian Fourier Projection maps the diffusion time t into a temporal embedding. Simultaneously, a dedicated pose encoder processes noisy samples into a 512-dimensional feature space. These components are concatenated and fed into a decoupled regression head—specifically an R_x, R_y, and T configuration—which employs MLPs initialized with a zero-module strategy to ensure stable training. Training is governed by a Variance Exploding (VE) SDE, where the network output is normalized by the noise scale $\sigma(t)$ to accurately estimate the score function. We use the Adam optimizer Kingma [2014] with an initial learning rate of 1×10^{-4} and apply an Exponential Moving Average (EMA) to model parameters to enhance stability. To increase efficiency, we implement a repeat training strategy that reuses point cloud features across multiple noise levels, using detached and cloned tensors to prevent computational graph conflicts during backpropagation.

A.2 Training and Inferring Details

Training Procedures. The training pipeline employs a dual-agent framework to optimize the generative and ranking components of the system. To enhance computational efficiency, we implement a data repetition strategy where geometric features are computed once and reused for multiple noise levels within a single batch. During this process, we explicitly decouple the point cloud features from the initial computational graph through detachment and cloning. This allows the network to perform multiple backward passes and gradient updates per batch without memory-related conflicts or unintended in-place modifications. Both the score-based and energy-based networks are optimized using the Adam algorithm with an initial learning rate of 1×10^{-4}, complemented by an exponential moving average (EMA) decay of 0.999 to stabilize long-term parameter convergence.

Inference and Temporal Tracking. At test time, 6D pose estimation is formulated as a continuous tracking task by predicting the relative transformation between consecutive frames. To generate candidates, we solve the Probability Flow ODE Song et al. [2020], sampling hypotheses from the learned pose posterior. To suppress temporal drift and error accumulation, we employ a recursive re-alignment strategy : the high-quality initial point cloud is transformed using the previous frame's estimated pose to serve as the reference geometry for the current observation. This ensures that the network always reasons from a deterministic historical prior even under severe visual data loss.

Ranking and Robust Aggregation. The generated candidates are ranked based on their approximated log-likelihoods. We implement a filtering threshold that discards the bottom 40% of low-confidence samples to eliminate potential outliers in the distribution. To resolve ambiguities arising from object symmetry or multi-finger occlusion, we apply a density-based clustering algorithm on the rotation manifold. The final rotation is aggregated from the dominant cluster by solving for the maximum eigenvector of the quaternion batch to strictly satisfy manifold constraints:

where represents the quaternions of the retained candidates within the densest cluster. The final translation is determined via mean pooling of the translational components within the same identified cluster.

A.3 SIM2REAL Gap Analysis

The sim-to-real gap in our method primarily stems from depth map quality. To address this, we apply a DPT-inspired depth refinement model to preprocess the noisy depth inputs from the L515 camera, which effectively bridges this gap. To further investigate the impact of depth noise, we train our model exclusively on simulated grasping and manipulation data and evaluate it directly on real-world data.

We employ Conditional Diffusion Models (CDMs) to perform denoising, completion, and edge enhancement on depth maps derived from the Real Grasp and Real Manip datasets. As illustrated in Figure 7, using the second frame of sequence carrot-seq_000 as a representative example, the figure displays the original RGB image (Top-Left), the raw input depth (Top-Right), the refined depth map (Bottom-Left), and the relative error map (Bottom-Right). A comparative analysis reveals that the

Table 6: Effect of depth refinement on sim-to-real transfer.

Depth Mode	REAL GRASP				REAL MANIP			
	$\text{ADD}_{\text{Err}} \downarrow$	$\text{ADDS}_{\text{Err}} \downarrow$	$\text{ADD}_{\text{AUC}} \uparrow$	$\text{ADDS}_{\text{AUC}} \uparrow$	$\text{ADD}_{\text{Err}} \downarrow$	$\text{ADDS}_{\text{Err}} \downarrow$	$\text{ADD}_{\text{AUC}} \uparrow$	$\text{ADDS}_{\text{AUC}} \uparrow$
SIM	0.0410	0.0216	59.02	78.46	0.0360	0.0186	53.85	77.45
SIM&REAL	**0.0342**	**0.0156**	**65.87**	**84.44**	**0.0334**	**0.0171**	**65.56**	**81.51**

refined depth map significantly mitigates noise artifacts Jiang et al. [2026] and reconstructs missing geometric information while preserving distinct object contours (e.g., the carrot). Consequently, the overall quality approximates that of ideal simulation-based depth maps, thereby facilitating robust downstream robotic manipulation tasks. It should be noted that pseudo-color mapping is applied here to enhance visual distinctiveness.

A.4 Experimental Details

We evaluate all baselines using their official implementations with standardized configurations to ensure fair comparison. Any6D utilizes the base_plus backbone with an input resolution of 1024×1024 and maintains 7 memory frames for temporal consistency. GigaPose is deployed with the large model variant, using an SO(3) grid size of 576 and performing 5 refinement iterations per pose hypothesis. Omni6DPose is configured with a diffusion temperature of $T_0 = 0.55$ during initialization (generating 50 pose candidates) and $T_0 = 0.15$ for tracking. FoundationPose executes 5 refinement iterations for initialization and reduces to 2 iterations during tracking, with a voxel size set to $1/20$ of the object diameter. Finally, OnePoseviaGen adopts the Hi3DGen pipeline (guidance strength 7.5/3) for all instances, operating on 518×518 input images. Crucially, for the scale recovery stage, we implement an adaptive reference image upscaling strategy: to prevent feature matching failures on small objects, the method iteratively tests magnification factors $\{1, 2, 4, 8, 16, 32\}$ to maximize SuperPoint correspondences.

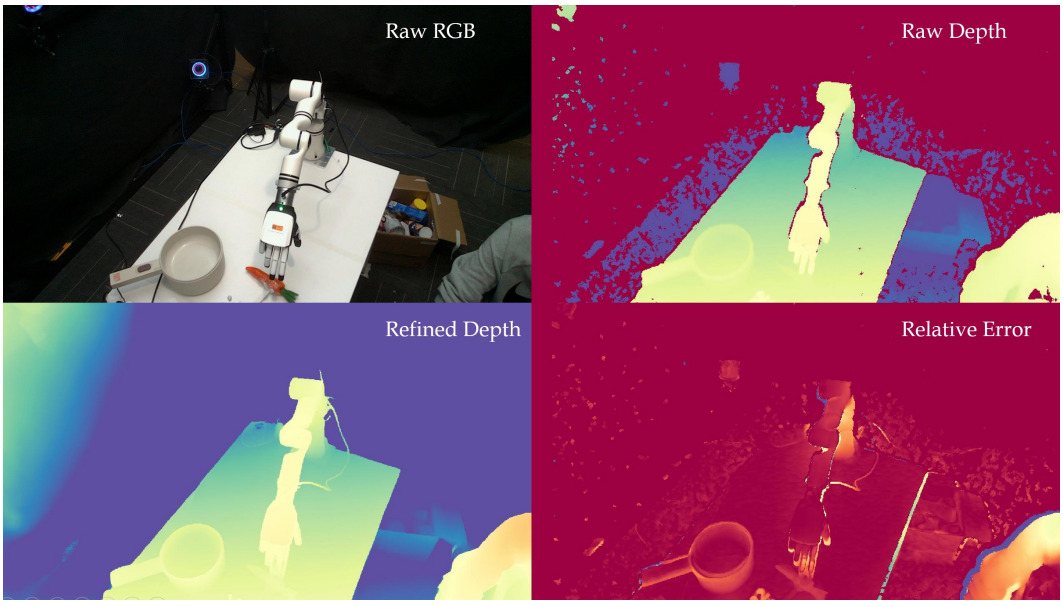

Figure 7: Visualization of Depth Refinement.

B Dataset Analysis

B.1 Frame Distribution across Object Categories

Figure 8 presents the statistical distribution of total data frames for each object category within the DexOPE dataset. With the dataset comprising a massive scale of 2 million frames, this statistic provides a granular view of the data volume available for training. As shown in the figure, the dataset covers a wide range of categories, from everyday objects to complex tools, with the frame count distribution reflecting the richness of interaction data collected. This extensive temporal coverage—averaging hundreds of frames per sequence—ensures that the dataset captures continuous dynamic transitions and diverse viewing angles, which is critical for learning robust, physics-aware 6D pose estimation models.

B.2 Statistics of Object Materials and Symmetries

In the domain of 6D object pose estimation, handling object symmetry and diverse material properties remains a principal challenge, particularly under the severe occlusions inherent to dexterous manipulation.

Object Materials

Real-world manipulation requires perception systems to handle diverse surface properties. DexOPE incorporates objects with varying optical characteristics, categorized into three main types: Diffuse, Specular, and Transparent. The left section of Figure 9 details the material distribution: the majority of objects are Diffuse (81.5%), while a significant portion are Specular (15.5%) or Transparent (3.0%). By including these challenging materials, DexOPE provides a rigorous benchmark for assessing the robustness of pose estimation models against reflections and transparency artifacts in realistic visual environments.

Object Symmetries

To rigorously evaluate pose estimation algorithms, DexOPE includes objects with a wide spectrum of symmetry attributes. As illustrated in the right section of Figure 9, we classify these objects into four symmetry types based on their geometric properties: Unimodal (asymmetric objects), Continuous (e.g., cylindrical bottles), and discretely symmetric objects classified as Bimodal or 24-peak. The statistical breakdown reveals that Unimodal objects constitute 52.2% of the dataset, followed by Continuous symmetries at 34.6%, Bimodal at 12.9%, and 24-peak symmetries at 0.3%. This distribution ensures that methods are tested against both unambiguous asymmetric targets and highly ambiguous symmetric geometries.

B.3 Visualization of DexOPE

To demonstrate the diversity and realism of our benchmark, Figure 10 presents selected visualization examples from the DexOPE dataset. The samples span all four data domains: Sim Grasp, Sim Manipulation, Real Grasp, and Real Manipulation. These visualizations highlight the high-fidelity rendering in simulation and the complex real-world setups involving the robotic hand and various objects. Notably, the samples showcase the severe visual occlusions caused by multi-finger interactions—a core challenge DexOPE aims to address. The dataset provides dense multimodal data, including RGB images, depth maps, and precise 6D pose annotations, ensuring a comprehensive platform for training and evaluating physics-aware pose estimation methods.

B.4 Comparison on Inference time.

C Specific Experiment Results

C.1 Detailed Experimental Results on Sim-Grasp

As shown in Table 8, our method demonstrates superior precision and robustness across diverse object configurations in the Sim-Grasp dataset, achieving the lowest mean ADD error of 0.0501 and ADDS error of 0.0248. While baselines like Any6D show competitive performance in specific metrics, our approach significantly outperforms them and other methods such as FoundationPose and

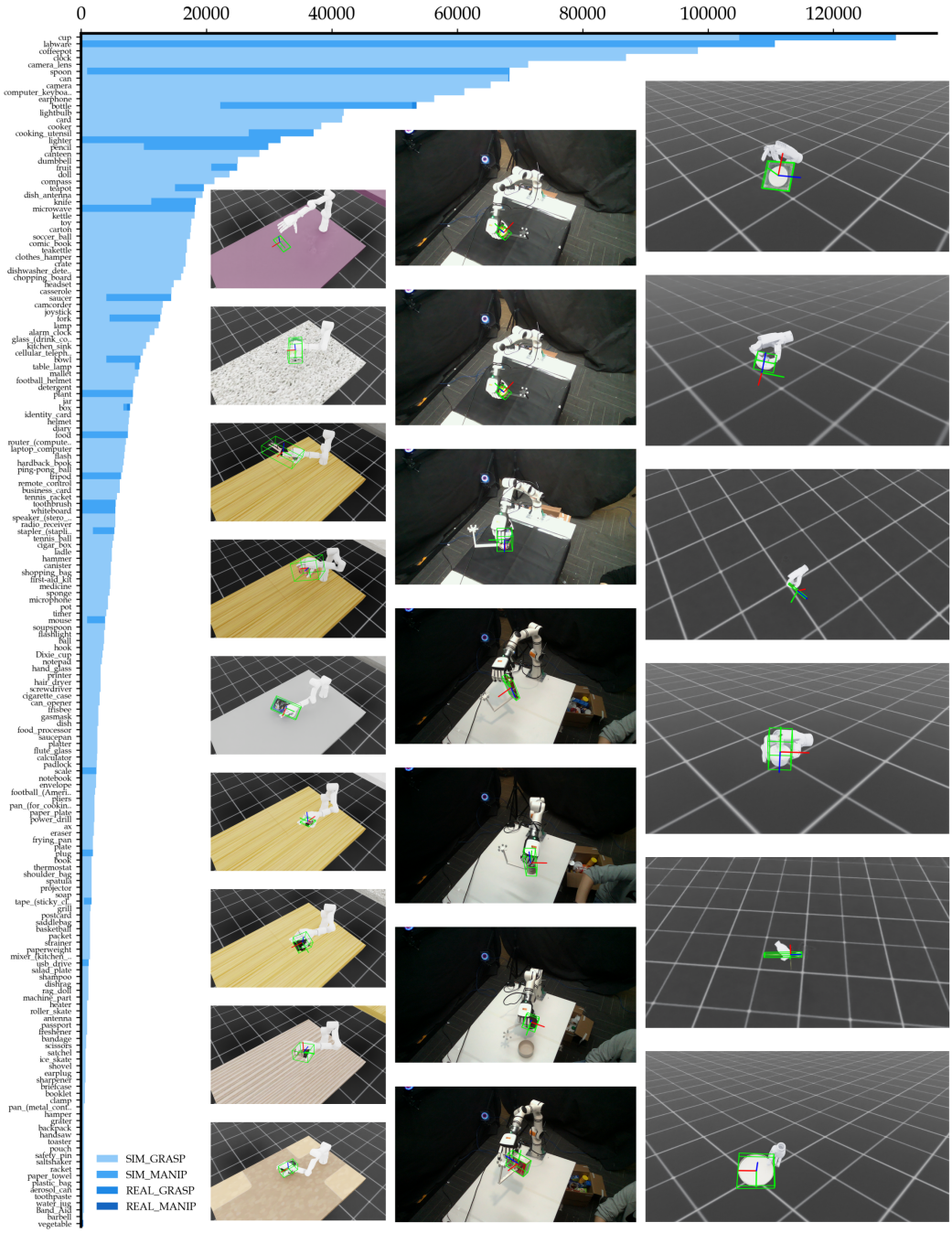

Figure 8: Frequency of occurrence for all object categories within DexOPE.

GigaPose, particularly in challenging scenarios like Task 052 where baseline errors spike drastically (e.g., Any6D reaches 0.2460); in contrast, our method maintains consistent accuracy with an overall AUC (ADDS) of 75.23, validating its reliability in static grasping environments.

C.2 Detailed Experimental Results on Sim-Manip

The results presented in Table 9 highlight the distinct advantage of our physics-guided framework in handling dynamic in-hand manipulation, where our method achieves a mean ADD of 0.0434,

significantly surpassing the runner-up Any6D at 0.1912. Unlike baselines such as Omni6DPose and GigaPose, which exhibit high volatility and frequent tracking failures in complex sequences like 06421@6 and 01857@10—with errors exceeding 0.30 in some cases—our model maintains stable pose estimation with a mean AUC (ADDS) of 74.14, proving its effectiveness in resolving ambiguities during continuous contact transitions.

C.3 Detailed Experimental Results on Real-Grasp

As demonstrated in Table 10, our method evidences strong sim-to-real generalization capabilities on the Real-Grasp dataset, recording the lowest mean ADD of 0.0342 and a dominant AUC (ADDS) of 84.44, whereas model-based approaches like GigaPose struggle with real-world noise, resulting in a much higher mean ADD of 0.3466. We observe that while Any6D and FoundationPose maintain reasonable tracking on simpler geometries, they falter on specific objects like bottle_rusuanjun and box_baichun, whereas our approach delivers consistent sub-centimeter accuracy across all test cases, outperforming the closest competitors by a substantial margin.

C.4 Detailed Experimental Results on Real-Manip

In the most challenging Real-Manip sequences characterized by severe occlusion and complex dynamics, Table 11 shows that our method achieves state-of-the-art performance with a mean ADD of 0.0334, which is an order of magnitude better than GigaPose's 0.4231. While baselines frequently

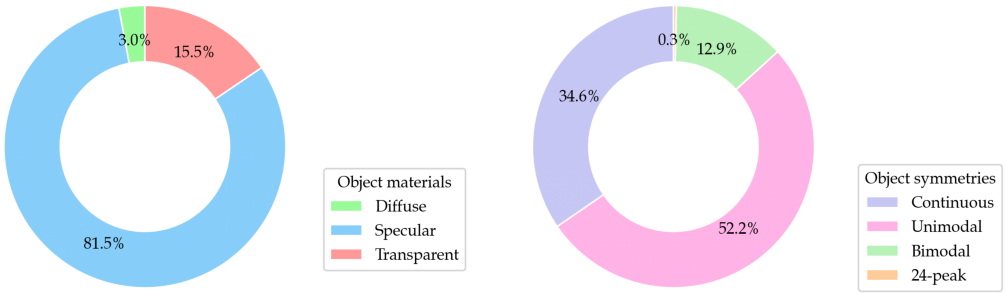

Figure 9: Statistics of Object Materials and Symmetries. Left: Materials are categorized into three classes: Transparent (based on opacity parameters d or Tr), Specular (objects with specular maps or high shininess $N_s > 10$), and Diffuse (matte surfaces with low specularity). Right: Symmetry types are classified based on geometric properties: Unimodal, Continuous, Bimodal, and 24-peak.

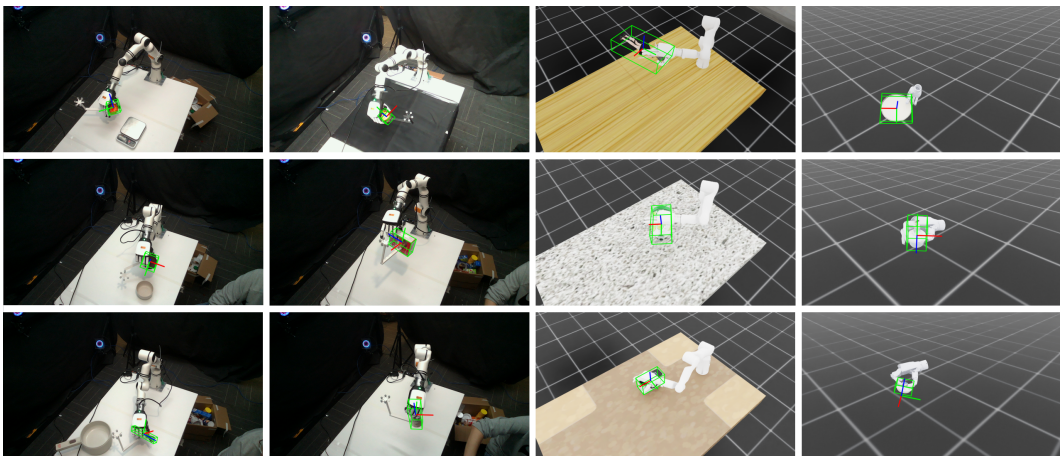

Figure 10: Visualization of DexOPE.

We have included an inference frame rate (FPS) comparison in our experiments. As shown in Table 7, our inference speed falls in the mid-range. Despite our lower FPS compared to some methods like CATRE, our approach still adequately fulfills real-world usage requirements, particularly for robot operations. Our primary focus is solving object pose estimation under severe occlusion during manipulation, and we plan to accelerate our deployment inference using TensorRT in future work. In terms of the sampling number, we generate initial noises and sample them in parallel. Therefore, this step does not incur additional latency.

Table 7: Inference speed comparison (FPS).

Method	**FPS** ↑
Any6D	17.0
GigaPose (Coarse+Refine)	12.3
Omni6DPose	13.0
FoundationPose	31.0
OneposeviaGen	0.77
Ours	11.5

lose track—evidenced by FoundationPose's complete failure on the cabbage task (0.00 AUC) and Any6D's significant error on the bottle_starbuck sequence—our system effectively utilizes tactile and physics guidance to maintain robust tracking, securing a mean AUC (ADDS) of 82.09 compared to the 74.56 of the next best method.

Table 8: Detailed per-task comparison on SIM GRASP sequences.

Method	Task ID	$\text{ADD}_{\text{Err}} \downarrow$	$\text{ADDS}_{\text{Err}} \downarrow$	$\text{ADD}_{\text{AUC}} \uparrow$	$\text{ADDS}_{\text{AUC}} \uparrow$
Any6D	001	0.0080	0.0045	92.04	95.41
	052	0.2460	0.2055	12.47	35.82
	035	0.0068	0.0039	93.26	96.24
	017	0.2083	0.1761	25.59	43.06
	009	0.0736	0.0294	29.69	70.66
	007	0.0197	0.0065	80.26	93.49
	Mean	0.0954	0.0710	**55.55**	72.45
GigaPose	001	0.1300	0.0901	43.30	51.08
	052	0.1456	0.1136	3.78	14.50
	035	0.1240	0.0966	10.09	23.71
	017	0.1419	0.1143	25.98	42.96
	009	0.1445	0.1281	19.81	28.44
	007	0.1284	0.0880	18.47	42.14
	Mean	0.1357	0.1051	20.24	33.81
Omni6DPose	001	0.1598	0.1241	48.96	55.63
	052	0.1692	0.1274	1.61	3.17
	035	0.1858	0.1468	2.03	11.26
	017	0.1978	0.1697	10.30	23.06
	009	0.1819	0.1598	21.96	31.74
	007	0.1684	0.1222	2.25	27.31
	Mean	0.1771	0.1406	14.52	25.36
Foundation Pose	001	0.0955	0.0622	58.47	61.18
	052	0.0584	0.0394	55.01	61.98
	035	0.0777	0.0529	51.07	59.80
	017	0.1225	0.0869	38.62	63.75
	009	0.1107	0.0836	58.77	62.53
	007	0.0656	0.0456	62.39	66.51
	Mean	0.0884	0.0618	54.06	62.63
OnePoseviaGen	001	0.1666	0.1037	14.37	51.86
	052	0.1406	0.0999	3.27	18.63
	035	0.0560	0.0218	45.62	78.48
	017	0.1653	0.1347	24.26	47.34
	009	0.1660	0.1158	4.97	20.04
	007	0.1036	0.0475	11.77	56.19
	Mean	0.1330	0.0872	17.38	45.42
Ours	001	0.0647	0.0382	35.3175	61.6883
	052	0.0329	0.0156	67.0940	84.3046
	035	0.0654	0.0272	34.5238	72.6912
	017	0.0379	0.0157	62.0464	84.3434
	009	0.0512	0.0311	48.6791	68.9200
	007	0.0487	0.0207	51.2065	79.4332
	Mean	**0.0501**	**0.0248**	49.81	**75.23**

Table 9: Detailed per-task comparison on SIM MANIP sequences.

Method	Task ID	$\text{ADD}_{\text{Err}} \downarrow$	$\text{ADDS}_{\text{Err}} \downarrow$	$\text{ADD}_{\text{AUC}} \uparrow$	$\text{ADDS}_{\text{AUC}} \uparrow$
Any6D	01857@10	0.0189	0.0078	81.15	92.15
	06421@6	0.4300	0.3855	27.57	29.55
	09200@0	0.0022	0.0020	97.92	98.08
	11cb7@0	0.3251	0.2863	19.53	21.63
	17ab7@0	0.3424	0.2939	11.18	22.11
	419a35	0.0286	0.0153	72.49	84.43
	Mean	0.1912	0.1651	51.64	58.00
GigaPose	01857@10	0.1958	0.1511	11.68	21.08
	06421@6	0.3370	0.2889	8.79	18.83
	09200@0	0.2206	0.1646	10.10	21.87
	11cb7@0	0.3836	0.3427	10.21	13.08
	17ab7@0	0.3300	0.2888	8.42	13.42
	419a35	0.1375	0.1131	13.35	17.36
	Mean	0.2674	0.2235	10.42	17.61
Omni6DPose	01857@10	0.5227	0.4806	6.96	8.88
	06421@6	0.0368	0.0134	63.15	86.63
	09200@0	0.4206	0.3710	16.86	21.76
	11cb7@0	0.1388	0.0947	10.43	16.41
	17ab7@0	0.2068	0.1572	4.56	6.02
	419a35	0.0631	0.0544	41.76	49.52
	Mean	0.2315	0.1952	23.95	31.54
Foundation Pose	01857@10	0.1015	0.0799	35.19	42.11
	06421@6	0.1444	0.1143	48.49	54.16
	09200@0	0.2107	0.1711	28.10	29.61
	11cb7@0	0.2681	0.2106	8.31	17.14
	17ab7@0	0.1717	0.1479	13.26	14.85
	419a35	0.0888	0.0078	12.87	92.47
	Mean	0.1642	0.1220	24.37	41.72
OneposeviaGen	01857@10	0.1147	0.0708	16.73	33.32
	06421@6	0.1882	0.1311	13.66	46.33
	09200@0	0.6743	0.6189	5.55	15.00
	11cb7@0	0.3734	0.3313	13.77	28.15
	17ab7@0	0.0829	0.0510	25.05	52.09
	419a35	0.0940	0.0492	14.57	50.81
	Mean	0.2546	0.2087	14.89	37.58
Ours	01857@10	0.0733	0.0447	32.4550	55.4018
	06421@6	0.0171	0.0110	82.7624	88.8669
	09200@0	0.0435	0.0222	56.5308	77.8474
	11cb7@0	0.0339	0.0173	66.0774	82.7020
	17ab7@0	0.0308	0.0233	69.1498	76.7256
	419a35	0.0616	0.0367	38.5101	63.2997
	Mean	**0.0434**	**0.0259**	**57.58**	**74.14**

Table 10: Detailed per-task comparison on REAL GRASP sequences.

Method	Task ID	$\mathrm{ADD}_{\mathrm{Err}} \downarrow$	$\mathrm{ADDS}_{\mathrm{Err}} \downarrow$	$\mathrm{ADD}_{\mathrm{AUC}} \uparrow$	$\mathrm{ADDS}_{\mathrm{AUC}} \uparrow$
Any6D	bottle_mmbeans	0.0450	0.0317	71.68	80.61
	bottle_muyulu	0.1288	0.0759	30.81	70.38
	bottle_rusuanjun	0.3549	0.3076	35.62	62.81
	box_baichun	0.1098	0.0597	34.29	73.50
	box_orio	0.0218	0.0073	78.28	92.73
	box_tango	0.0180	0.0113	85.48	90.74
	box_wangwang	0.2938	0.2662	52.40	60.95
	shupian_heisonglu	0.0848	0.0439	46.99	82.78
	paomian_haixian	0.1351	0.1141	64.73	74.75
	Mean	0.1324	0.1020	55.59	76.58
GigaPose	bottle_mmbeans	0.3805	0.3401	11.93	24.18
	bottle_muyulu	0.3434	0.2771	7.92	16.04
	bottle_rusuanjun	0.4346	0.3596	4.92	10.51
	box_baichun	0.2809	0.2220	10.06	24.17
	box_orio	0.3839	0.2914	8.96	19.60
	box_tango	0.4062	0.3336	3.05	8.36
	box_wangwang	0.2779	0.2221	15.62	28.83
	shupian_heisonglu	0.3567	0.2779	5.57	18.11
	paomian_haixian	0.2553	0.2040	13.04	32.82
	Mean	0.3466	0.2809	9.01	20.29
Omni6DPose	bottle_mmbeans	0.0903	0.0665	43.71	55.36
	bottle_muyulu	0.0583	0.0391	60.94	70.67
	bottle_rusuanjun	0.1785	0.1424	30.54	38.73
	box_baichun	0.1200	0.0916	53.15	63.50
	box_orio	0.0632	0.0354	36.72	64.55
	box_tango	0.0325	0.0205	71.62	80.82
	box_wangwang	0.1087	0.0797	38.11	50.77
	shupian_heisonglu	0.1006	0.0697	52.50	60.46
	paomian_haixian	0.0816	0.0574	53.31	63.78
	Mean	0.0926	0.0669	48.96	60.96
Foundation Pose	bottle_mmbeans	0.2069	0.1664	29.45	55.63
	bottle_muyulu	0.0823	0.0379	32.50	65.71
	bottle_rusuanjun	0.1836	0.1126	12.13	40.71
	box_baichun	0.0995	0.0302	20.94	71.37
	box_orio	0.1066	0.0906	58.21	59.98
	box_tango	0.0535	0.0261	56.27	73.93
	box_wangwang	0.0531	0.0277	54.67	72.77
	shupian_heisonglu	0.0336	0.0157	69.66	86.47
	paomian_haixian	0.0814	0.0253	27.47	74.78
	Mean	0.1001	0.0592	40.14	66.82
OnePoseviaGen	bottle_mmbeans	0.1337	0.0918	26.18	53.80
	bottle_muyulu	0.1796	0.1228	10.55	49.03
	bottle_rusuanjun	0.4359	0.3673	8.20	27.02
	box_baichun	0.5733	0.5175	9.77	32.38
	box_orio	0.1134	0.0526	12.88	47.83
	box_tango	0.1685	0.1101	11.59	41.45
	box_wangwang	0.1947	0.1431	16.20	42.61
	shupian_heisonglu	0.1534	0.0853	8.99	40.66
	paomian_haixian	0.1370	0.0841	21.60	54.59
	Mean	0.2322	0.1750	14.00	43.26
Ours	bottle_mmbeans	0.0415	0.0211	58.48	78.97
	bottle_muyulu	0.0308	0.0146	69.10	85.46
	bottle_rusuanjun	0.0300	0.0138	69.97	86.23
	box_baichun	0.0323	0.0146	67.66	85.40
	box_orio	0.0453	0.0169	54.71	83.16
	box_tango	0.0344	0.0160	65.67	83.97
	box_wangwang	0.0330	0.0140	66.87	85.95
	shupian_heisonglu	0.0298	0.0159	70.71	84.19
	paomian_haixian	0.0303	0.0134	69.68	86.62
	Mean	**0.0342**	**0.0156**	**65.87**	**84.44**

Table 11: Detailed per-task comparison on REAL MANIP sequences.

Method	Task ID	$\mathrm{ADD}_{\mathrm{Err}} \downarrow$	$\mathrm{ADDS}_{\mathrm{Err}} \downarrow$	$\mathrm{ADD}_{\mathrm{AUC}} \uparrow$	$\mathrm{ADDS}_{\mathrm{AUC}} \uparrow$
Any6D	bottle_starbuck_step	0.3542	0.3033	24.57	57.60
	cabbage	0.0209	0.0117	85.00	92.07
	carrot	0.0193	0.0079	80.75	92.13
	huanggua_basket	0.0227	0.0072	77.24	92.85
	nangua	0.0209	0.0074	79.17	92.56
	pour	0.3278	0.2848	16.76	20.15
	Mean	0.1276	0.1037	60.58	74.56
GigaPose	bottle_starbuck_step	0.3969	0.3408	10.14	15.43
	cabbage	0.3522	0.2245	1.69	6.01
	carrot	0.4050	0.3190	6.39	20.03
	huanggua_basket	0.7647	0.6704	2.76	2.76
	nangua	0.1840	0.1571	35.79	48.59
	pour	0.4358	0.3805	5.29	14.59
	Mean	0.4231	0.3487	10.34	17.90
Omni6DPose	bottle_starbuck_step	0.2018	0.1578	36.48	40.55
	cabbage	0.1134	0.0882	57.46	66.89
	carrot	0.0364	0.0255	70.19	78.67
	huanggua_basket	0.0286	0.0178	71.41	82.27
	nangua	0.0994	0.0837	54.69	59.24
	pour	0.3509	0.2967	8.37	12.54
	Mean	0.1384	0.1116	49.77	56.69
FoundationPose	bottle_starbuck_step	0.0871	0.0199	13.01	80.06
	cabbage	0.2607	0.1243	0.00	53.17
	carrot	0.0413	0.0216	58.96	78.61
	huanggua_basket	0.1551	0.0073	0.00	92.61
	nangua	0.0518	0.0350	49.66	64.99
	pour	0.0489	0.0276	51.87	72.34
	Mean	0.1075	0.0393	28.92	73.63
OnePoseviaGen	bottle_starbuck_step	0.1479	0.0966	4.43	21.29
	cabbage	0.1135	0.0506	24.95	61.00
	carrot	0.1406	0.0951	18.00	44.08
	huanggua_basket	0.1027	0.0300	20.93	69.92
	nangua	0.1390	0.1089	41.78	60.06
	pour	0.2683	0.2109	3.50	15.31
	Mean	0.1520	0.0987	18.93	45.28
Ours	bottle_starbuck_step	0.0311	0.0146	68.92	85.45
	cabbage	0.0204	0.0109	64.89	78.60
	carrot	0.0243	0.0161	75.63	84.01
	huanggua_basket	0.0145	0.0085	85.64	91.70
	nangua	0.0311	0.0181	72.93	83.62
	pour	0.0790	0.0342	25.33	65.68
	Mean	**0.0334**	**0.0171**	**65.56**	**81.51**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state our three main contributions: the large-scale dataset (1.5k objects, 2M frames), the physics-guided visual-tactile diffusion framework, and the experimental results demonstrating state-of-the-art performance. See Section 1 and Abstract.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 5, including the requirement of an initial unoccluded frame and the inference speed falling short of real-time performance.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include formal theoretical proofs. The methodology is based on score-based diffusion models with empirical validation.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Architecture details, training procedures, hyperparameters, and inference details are provided in Appendix A. Baseline configurations are described in Appendix A.4.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The dataset and code are not publicly released at the time of submission.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training details including optimizer (Adam, lr=1e-4), EMA decay, and inference procedures are provided in Appendix A.1 and A.2. Baseline configurations are in Appendix A.4.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported due to the high computational cost of running multiple trials across the large-scale dataset. Results are reported as mean values across all test sequences.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: Detailed compute resource information is not included in the current version.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

   Answer: [Yes]

   Justification: The research involves robotic manipulation and does not raise ethical concerns. Data collection via teleoperation involved only the authors and no human subjects were harmed.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational research on robotic perception. The primary application is enabling dexterous robot manipulation, which has broad positive implications. No significant negative societal impacts are foreseen.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse?

    Answer: [N/A]

    Justification: The dataset and models are focused on robotic manipulation and pose estimation, posing no significant risk for misuse.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets used in this paper (YCB, Objaverse, OakInk2, ManipTrans, FoundationPose, etc.) are properly cited in the references.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The DexOPE dataset is documented in Section 3.1 and Appendix B, including statistics of object categories, materials, symmetries, and data collection procedures across four domains.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing. Teleoperation data was collected by the authors themselves using a MANUS VR glove.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects research requiring IRB approval. Data collection was performed by the research team using robotic systems.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

Answer: [N/A]

Justification: LLMs are not used as part of the core methodology.